Land Cover Mapping in Support of the National Park Service Vegetation Mapping Program Using Multi-temporal Landsat 7 Data and a Decision Tree Classifier

Eric C. Brown de Colstoun[1], Michael H. Story[2], Craig Thompson[3], Kathy Commisso[3], Timothy G. Smith[3], and James R. Irons[4]

[1] Science Systems and Applications, Inc., Code 923, NASA/Goddard Space Flight Center, Greenbelt, MD 20771 USA. Tel.: (301) 614-6597; Fax: (301) 614-6695; e-mail: ericbdc@ltpmail.gsfc.nasa.gov
[2] Natural Resource Information Division, National Park Service, Denver, CO USA.
[3] Delaware Water Gap National Recreation Area, National Park Service, Milford, PA USA.
[4] Biospheric Sciences Branch, NASA/Goddard Space Flight Center, Greenbelt, MD USA.

ABSTRACT

Decision tree classifiers have received much recent attention, particularly with regards to land cover classifications at continental to global scales. Despite their many benefits and general flexibility, the use of decision trees with high spatial resolution data has not yet been fully explored. In support of the National Park Service Vegetation Mapping Program, we have examined the feasibility of using a commercially available decision tree classifier with multi-temporal satellite data from the Enhanced Thematic Mapper-Plus (ETM+) instrument to map 11 land cover types at the Delaware Water Gap National Recreation Area near Milford, PA. Ensemble techniques such as boosting and consensus filtering of the training data were used to improve both the quality of the input training data as well as the final products.

Using land cover classes as specified by the National Vegetation Classification Standard at the Formation level, the final land cover map has an overall accuracy of 82% (Kappa=0.80) when tested against a validation data set acquired on the ground (n=195). This same accuracy is 99.5% when considering only forest vs. non-forest classes. Usage of ETM+ scenes acquired at multiple

dates improves the accuracy over the use of a single date, particularly for the different forest types. These results demonstrate the potential applicability and usability of such an approach to the entire National Park system, and to high spatial resolution land cover and forest mapping applications in general.

I. INTRODUCTION

Traditional approaches to land cover classification from remotely-sensed data have typically relied on statistical classifiers such as supervised Maximum Likelihood Classifiers (MLC) or unsupervised isoclustering techniques, to name but a few (e.g. Swain and Davis 1978; Richards 1983). Increasingly, advances in the fields of pattern recognition and machine learning have led to the application of decision tree and neural network classifiers, particularly with regards to land cover classifications at global to continental scales (DeFries et al. 1998; Strahler et al. 1999; Hansen et al. 2000). In fact, decision trees are also used in global land cover classification algorithms for the MODerate Resolution Imaging Spectroradiometer (MODIS) (Strahler et al. 1999) and are planned for use with data from future instruments such as the Visible/Infrared Imager/Radiometer Suite (VIIRS) to be flown onboard the National Polar-orbiting Operational Environmental Satellite System (NPOESS) (Brown de Colstoun et al. 2000).

Decision trees have been preferred to statistical classifiers for these coarse scale applications because they do not make any implicit assumptions about normal distributions in the input data, as an MLC would (Hansen et al. 1996; Friedl and Brodley 1997). These classifiers can also accept a wide variety of input data, including non-remotely-sensed ancillary data, and in the form of both continuous and/or categorical variables. The general simplicity and hierarchical structure of the results from decision trees can also be valuable assets to both experienced and inexperienced users

for interpretation, algorithm testing and refinement, and analysis. Finally, decision trees have been shown to provide improved accuracies over the use of other more traditional classifiers (e.g. Strahler et al. 1999; Brown de Colstoun et al. 2000). In spite of these proven benefits, however, the use of decision trees for applications with high spatial resolution data such as Landsat Thematic Mapper (TM) or Enhanced Thematic Mapper-Plus (ETM+) has not yet been fully explored.

The goal of the research reported here is to evaluate the performance of the commercially-available decision tree classifier C5.0 (Quinlan 1993) for land cover classifications using 30 m resolution data from the ETM+ instrument onboard the Landsat 7 satellite. During this evaluation, methods developed primarily with coarse scale satellite data were tested with the ETM+ data, namely the use of vegetation phenology information (DeFries et al. 1995), or data acquired during different seasons, consensus filtering of training data (Brodley and Friedl 1997), and ensemble classifier techniques such as boosting (e.g. Friedl et al. 1999; DeFries and Chan 2000). The overarching goal of the research is to develop a robust, simple and easily repeatable methodology for classification that addresses the specific needs of the Vegetation Mapping Program (VMP), jointly administered by the National Park Service (NPS) and the Biological Resources Division of the United States Geological Survey (USGS).

II. BACKGROUND

A. Decision Trees Techniques and Remote Sensing

The use of decision trees as a viable alternative to more traditional classifiers has been explored primarily within the context of global or continental scale land cover classifications (DeFries et al. 1998; Hansen et al. 1996, 2000; Strahler et al. 1999; Friedl and Brodley 1997; Muchoney et al. 2000). The majority of these studies have utilized data acquired by the Advanced

Very High Resolution Radiometer (AVHRR) instrument at fairly coarse spatial scales ranging from 1 degree to 1km. Typically, these studies have used AVHRR data acquired over an entire year and have used the temporal evolution of either the Normalized Difference Vegetation Index (NDVI) and/or the individual spectral bands of the AVHRR as attributes for the classification. DeFries et al. (1995; 1998) and Hansen et al. (2000) derived temporal metrics from a full year of AVHRR data and exploited the temporal changes in reflectance/NDVI/brightness temperatures to successfully discriminate global land cover types.

The use of Landsat scenes acquired at different seasons and/or years to improve land cover classifications is certainly not a new concept. Among several other studies, Wolter et al. (1995) used multiple scenes of Landsat Multispectral Scanner (MSS) and TM for five years to improve the mapping of forest types in northern Wisconsin. They also provided an extensive review of the literature in using multi-temporal Landsat data for forestry applications. Pax-Lenney and Woodcock (1997) used multiple Landsat scenes acquired during the same year as well as different years to examine the status of agricultural lands in the Nile Delta in Egypt. Finally, Landsat TM scenes acquired during multiple seasons but the same year, so called leaf on/leaf off images, form the basis for regional land cover classifications being developed by the USGS within the context of a national land cover map of the United States (Vogelmann et al. 1998; Zhu et al. 2000). Clearly, the use of temporal information for classifications provides benefits at both fine and coarse resolutions. These benefits can easily be explored with decision tree classifiers, as they can provide indications of the relative importance of particular attributes or dates for certain land cover classes, and their interactions, or be used as a tool for data reduction or feature space exploration (Hansen et al. 1996).

As has been shown in Hansen et al. (1996) and DeFries et al. (1998), training data, and particularly global scale training data, tend towards non-gaussian distributions in spectral and/or

temporal feature space, implying a need for alternate methods to classification beyond parametric classifiers. Although the non-gaussian behavior of training data may be more severe at coarser spatial scales, it is also encountered at finer scales, if one considers a common bare soil category that may contain soils of varying brightnesses, for example. Decision trees and neural networks are well suited to this type of problem because they are non-parametric, in that they do not make any implicit assumptions about normal distributions in the input data. Both algorithms tend to produce comparable classification accuracies when tested with the same remotely-sensed data inputs (Strahler et al. 1999) and typically outperform other classifiers in terms of classification accuracies (e.g. Hansen et al. 1996; Strahler et al. 1999). However, decision trees are typically less computationally expensive than most neural networks (Weiss and Kulikowski 1991) and, by virtue of their hierarchical structure, also provide analysts and users with a simpler yet robust method to interpret, test, and analyze their results (Hansen et al. 1996; Friedl and Brodley 1997).

The flexible nature of decision trees and the general availability of commercial decision tree classifiers such as C4.5 (Quinlan 1993), and its successor C5.0, have aided recent advances in the field of machine learning, ensemble classifiers and consensus filtering of the training data being just two of these advances. A series of classifiers such as decision trees, termed an ensemble, can be combined to produce higher classification accuracies than any one of the particular classifiers. This reasoning is the foundation for ensemble classifier methods such as boosting and bagging (Quinlan 1996; Freund and Schapire 1997). In a boosted tree, a series of decision trees is created in an iterative fashion, with each successive tree focusing on the errors of the previous tree, or those instances that are most difficult to classify (Friedl et al. 1999). A boosted tree is then produced by voting amongst the different trees that have been created. Boosting has been shown to produce improved results over standard decision trees (Quinlan 1996; Friedl et al. 1999; DeFries and Chan 2000). DeFries and Chan

(2000) also indicate that boosted trees are more resistant to both random noise in the input data attributes and errors in the training data introduced by mislabeling during the training phase.

In consensus or majority filtering of the training data, random blocks of training/testing samples are drawn from the original training data a number of times. Each training set is used to produce a decision tree, with only those instances that have been classified incorrectly by all or a majority of the trees being removed from the training data. The training data generated from such an exercise are 'cleaned' or optimal in the sense that they minimize the presence of potentially mislabeled training instances and noisy data (Brodley and Friedl 1997). The underlying assumption of these techniques is that mislabeling errors are almost always introduced in the training process. This is particularly true when one considers the development of training data at coarse scales of one to several kilometers but is also true when training data are produced using high spatial resolution data such as TM or ETM+. While great care is taken to ensure that the training data describe the particular land cover type well, the process of generalizing multiple point observations to a polygon such as a field inevitably introduces some errors in the training data. The usage of a consensus filter to minimize these errors has been shown to be conservative with regards to both identifying the mislabeled training data, and not discarding training data that are accurate if perhaps different (Brodley and Friedl 1997). Brodley and Friedl (1997) show that the consensus filtering approach can also be implemented using different classifiers besides just decision trees and generally improves classification accuracies for a global data set where they have introduced artificial training errors at various levels. In this study we aim to capitalize on all of the 'lessons learned' using decision trees for coarse scale land cover applications, but here we examine the performance of these techniques when using high spatial resolution data from the ETM+ instrument. The study is also made in the context of the NPS/USGS Vegetation Mapping Program, described below.

B. The NPS/USGS Vegetation Mapping Program

The National Park Service initiated in 1992 the Vegetation Mapping component of its Inventory and Monitoring Program. The Inventory and Monitoring Program is charged with developing methodologies and protocols for the systematic accounting of all the natural resources of over 270 park units managed by the NPS, including land cover. The goal of the vegetation mapping component of the I&M Program was to develop a uniform hierarchical vegetation classification standard and methodology and to apply that standard and methodology to generate vegetation maps for most of the NPS park units. This includes sampling the vegetation, developing local classifications, describing the associations, and mapping the spatial extent of the vegetation associations. The vegetation maps are intended to support a wide variety of resource assessment, management, and conservation concerns at the park unit, regional, and national levels.

The vegetation mapping component of the I&M Program has established standards for the content, scale, accuracy, and format of the map products resulting from the application of the protocols. The vegetation classification scheme used by all of these projects follows the National Vegetation Classification Standard (NVCS) specified by the Vegetation Subcommittee of the Federal Geographic Data Committee (FGDC) (FGDC, 1997). The goal in developing the NVCS was to provide a classification standard that facilitates the interchange and comparison of vegetation information produced by different national agencies and organizations working at a range of scales and on a wide variety of resource management issues. The NVCS is a hierarchical classification system based upon salient physiognomic and floristic characteristics of terrestrial vegetation (FGDC, 1997). The Standard provides a list of mutually exclusive vegetation categories within each of nine hierarchical levels. The levels, from most general to most detailed, are: Division, Order, Physiognomic Class, Physiognomic Subclass, Physiognomic Group, Subgroup, Formation, Alliance,

and Association. The middle five levels, Physiognomic Class through Formation, are based on the physical structure of the vegetation (physiognomy). The two most detailed levels, Alliance and Association, are based on species composition (floristics). The system is also designed to allow aggregation of categories at one level into the more general categories at the next level in the hierarchy.

The mapping protocol used in these projects principally employs aerial photography, manually interpreted in conjunction with ancillary field data. Independent accuracy assessment based on field observations constitutes a critical component of the protocol. The resulting vegetation maps are required to provide a classification accuracy of at least 80% per class, a minimum mapping unit of 0.5 hectare at a scale of 1:24,000. The amount of time required to produce these highly detailed maps for an individual park unit typically amounts to two to five years depending on the unit size and complexity of vegetation species associations present.

The vegetation mapping component of the I&M Program was originally begun as a 10-year program. Vegetation maps have been completed for 18 park units since 1994. An additional 10 parks are near completion and mapping is in progress in another 58 units. Figure 1 shows the current status of the vegetation mapping component of the I&M Program within the conterminous U.S. (see http://biology.usgs.gov/npsveg/ for more details). While this effort is producing high quality, detailed maps, many parks require more timely information on land cover than is currently available. The mapping approach presented here is entirely consistent with the mapping methodology developed by the vegetation mapping component of the I&M Program, following the NVCS, the protocols for mapping and accuracy assessment, and the standards for final products. This approach involves the application of the C5.0 decision tree algorithm, using boosting and consensus filtering of training data, to the classification of terrain-corrected ETM+ data using imagery acquired at two dates

(Fall/Spring). The final products that will be derived from the ETM+ data will be less detailed than the maps currently derived from aerial photography, but will be produced more rapidly and at a lower cost, addressing the immediate needs of many parks for current land cover information. While the use of Landsat data may be expected to yield greater benefits for larger parks, the spatial coverage can allow both small and large parks to be placed and studied within a more regional context. The spatial coverage provided by Landsat may also allow the imaging of multiple park units within one scene, and the repeated coverage over time may enable more active monitoring of these park lands than would be currently available. Results from a pilot project at Delaware Water Gap National Recreation Area (NRA) are given here.

III. DATA AND METHODS

A. Study Area

The Delaware Water Gap National Recreation Area is located between the towns of Matamoras, PA at the northern terminus and Delaware Water Gap, PA to the south, overlapping the state line between the states of Pennsylvania and New Jersey. The park covers an area of over 27,000 hectares, roughly between latitudes 41°21'N and 40°26'N and longitudes 75°10'W and 74°44'W (see Figure 2). The vegetation of the park is dominated principally by broadleaf deciduous forests with a mix of different species, with some interspersed areas of evergreen needleleaf and mixed evergreen/deciduous forests. Of particular interest to park managers is the health of the hemlock forests that typically grow on steep ravines near creeks and streams. An aphid-like insect called the Hemlock Woolly Adelgid has devastated large stands of these unique forests.

Because of its historical heritage, many areas within the park contain previous agricultural areas in various stages of regrowth. However, many of these agricultural areas still remain active

today under park management. Important areas of cropland skirt the eastern border of the park along with some smaller urban areas to the SW and NE and more recent low-density residential developments to the West of the park. The elevation of the park ranges from about 80 m at the Delaware River to 488 m along Kittatiny Mountain which run SW-NE on the eastern boundary of the park. Immediate areas surrounding the park can reach altitudes of around 640 m.

B. ETM+ Data Processing

In order to capitalize on temporal differences in the signal of the different cover types, two pairs of cloud-free, terrain-corrected Landsat 7 ETM+ scenes for two dates (September 23, 1999 and January 29,2000) were acquired from the EROS Data Center (EDC) in Sioux Falls, SD. The scene for September was acquired before senescence of the deciduous vegetation and the January scene contained a substantial amount of snow cover. The scenes covered Path/Row 14/31 and 14/32 in the Landsat Worldwide Reference System and were delivered registered to a Universal Transverse Mercator (UTM) projection using a North American Datum 1983, as specified by VMP protocols. The two scenes for each date were mosaiced together to provide complete coverage of the park. In consultation with park scientists and staff, a subset of these scenes covering the park and all of the watersheds draining into the park were extracted for classification. ETM+ bands 2, 3, 4, 5, and 7 were used for the classification, along with the Normalized Difference Vegetation Index (NDVI), calculated as the ratio of the difference in digital counts of bands 4 and 3 divided by their sum. In this research, the atmosphere was assumed to be constant over the park and surrounding areas for each date.

C. Field Visits

Training and validation data were acquired mostly within the park during separate field visits in September 2000 and August 2001. During these visits digital photographs were acquired at each

site and were tagged with locational information acquired simultaneously with a Global Positioning System (GPS). The GPS camera system has a real-time differential GPS receiver system attached to a digital camera. The camera runs a script that queries the GPS unit for the time/data and position. The script will use that information to place a "watermark" on to a digital photo. Additionally, the script produces a comma delimited ASCII text file that can be downloaded from the camera with all of the information, such as file name, date, time, position, etc. These photographs are particularly useful for training and evaluation because they also provide a spatial context to the site visited and the cover types that are present in its general vicinity (Figure 3). The photos also become part of the record for any particular park and can assist the task of monitoring by repeat visits to the same site over time.

The training visits served to delineate areas around each study site within the ETM+ images where a particular cover type was known to exist with high confidence. Other training areas in the areas surrounding the park were also delineated based on similarity of spectral/temporal changes in the ETM+ digital counts. A total of 13449 training pixels were delineated in this fashion on the imagery. Because sufficient samples were not obtained for either needleleaf or deciduous open tree canopy (i.e. Woodlands) and Shrublands categories, only single Woodlands and Shrublands classes were considered. An Urban/developed category was added to split the NVCS non-vegetated category in two, again based on the needs of the park scientists. Table 1 shows the land cover classes considered and the number of pixels used for each. A random sampling strategy stratified by land cover type was used in the evaluation or validation phases. Some 140 sites were visited during this phase. An additional 57 points were extracted visually from the imagery and air photography and/or crops/managed fields GIS coverages of the park.

D. Description of the C5.0 Decision Tree

In their simplest form, decision tree classifiers successively partition the input data into more and more homogeneous subsets by producing optimal rules or decisions, also called nodes, which maximize the information gained and thus minimize the error rates in the branches of the tree (Weiss and Kulikowski 1991). Each final leaf is then the result of following a set of mutually exclusive decision rules down the tree.

The C4.5 and C5.0 decision trees are described in detail in Quinlan (1993), Friedl and Brodley (1997) and DeFries and Chan (2000). The C5.0 and C4.5 decision trees use the gain ratio to determine both the best attribute to separate the different classes in the training data as well as the best possible threshold to make this separation (Quinlan 1993). This process of recursively dividing the training data into smaller and smaller subsets continues until the leaves of the tree contain only cases from one class or until the splitting does not provide any improvement in the gain ratio.

A decision tree created in this fashion typically "overfits the data" and is generally quite accurate with respect to the training data that were used to produce it. However, because the training data typically contain labeling errors as well as attribute noise, some of the splits will be created from this noisy data and may in fact give very poor results when applied to unseen data or new cases (Weiss and Kulikowski 1991; Quinlan 1993). Such a tree will thus provide little predictive capability which is in fact one of the goals of building a decision tree in the first place. At this point the tree is generally pruned back starting at the leaves and moving upward by considering whether the information gain from one sub-tree, and the whole tree by extension, can be improved by replacing it with either its most common leaf or a branch. This process is fully automated in C5.0 and does not require a separate sample of the training data, as in other tree pruning methods (e.g. Breiman et al. 1984). The severity of pruning in C5.0 can be adjusted in two ways. In the first, the user can specify

the minimum number of cases that must follow each of the branches of a tree. The second is based on a user-specified confidence level used to calculate the predicted error rate at each leaf, branch and/or sub-tree, as well as the predicted number of errors for each of these. Smaller values create more severe pruning. A pruned tree created by adjusting these two parameters results in a decision tree that is smaller and more generalized, does not perform as well on the original training data, but on the other hand can provide improved accuracy when applied to new cases (Quinlan 1993, 1999).

C5.0 also incorporates new methods in machine learning such as boosting. In boosting, an initial decision tree is created as described above and forms the basis for building a series of subsequent trees that are forced to focus on the errors of the previous one. This procedure is performed by updating a set of weights assigned to each training case and which are set equal in the initial tree or trial. After the first tree is created those weights for erroneous cases are proportionally increased, and those for correct cases are proportionally decreased. The next tree is produced with these new weights and the process is repeated, classifying in essence the "difficult' cases found in the training data (Schapire 1999). Voting amongst the different trees that have been created in this fashion then produces a boosted tree. After 10 such trials on 27 different data sets, Quinlan (1996) shows that boosting reduces the amount of errors by about 15% on average over the use of a single tree.

For this study, the 13449 training samples were randomly divided into 20 equal-sized training and testing blocks. The C5.0 tree was run for each of the 20 training blocks using fairly severe pruning in order to produce smaller, more generalized trees that would better capture any mislabeled training data. These 20 trees were then applied to the unseen testing blocks and those training/testing pixels misclassified by all 20 trees were discarded, following Brodley and Friedl (1997). The remaining training pixels were then input into a boosted tree to produce the final decision trees. These

boosted trees were applied to the entire subset of Landsat data to produce the final products. The use of one Landsat scene versus multiple scenes was evaluated by considering the overall and per-class classification accuracies obtained with either or both of the scenes on the 50% testing samples. In a preliminary step, elevation, slope and aspect information were planned to be used in conjunction with the Landsat data but the training data for each class were found across a broad spectrum of elevation, slope and aspects. The performance of C5.0 using only the Landsat 7 attribute data and the training data was also internally evaluated using the 50% testing samples that are kept separate from the tree-building process. Finally, labels assigned by C5.0 in the final tree were compared to those determined from field visits to provide an independent accuracy assessment of the land cover map, following methods described in Congalton (1991).

IV. RESULTS AND DISCUSSION

The mean accuracy obtained on the 20 50% testing blocks can be used for internal evaluation of the performance of the classifier with these training data. Because the training and testing data in this case are not truly independent, these results may typically provide an optimistic estimate when compared to results obtained against independent sources. We also use this internal evaluation to ascertain the effect of using one or multiple Landsat scenes for the classification. When using only the January scene, the mean error rate on the 20 testing samples (n=6727) is 30.01% (±0.32%). In this case there is some substantial confusion between the three forest types and the Woodlands category but, more significantly, nearly all short vegetation types as well as bare categories are severely confused, including the water bodies that are covered with snow and/or ice. In contrast, the results obtained for the three forest types are quite good (e.g. Evergreen Needleleaf Forests have producer's and user's accuracies of 96.81% (±0.67%) and 94.67% (±0.58%), respectively). When using only the

September 23 scene, the average classification error was 28.49% (±0.42%) over the same testing samples, somewhat better than using only the January scene. Using only this scene, the results for the forest types show moderate to good producer's and user's accuracies, ranging from 46.74% to 83.16% and 54.53% to 86.77%, respectively, from the Mixed Forest class to the Evergreen Needleleaf Forests class. The Deciduous Broadleaf Forest category provided intermediate values from the other two forest classes. The accuracies for all short vegetation types increase substantially when compared to those obtained with the January scene. Classes 7, 9, 10, and 11 all have users and producers accuracies of 80% or better when using only the September scene, with Classes 6 and 8 with values above 70%.

When comparing the mean user's and producer's accuracies obtained with single scenes for either date, the separation of all the forest types is decreased from January to September, with differences of over 30% in some cases such as the Mixed Forest class producer's accuracies. In contrast, other classes such as wetlands show large improvements of over 70% in both users and producers accuracies when using the September data over the January data. More moderate but still substantial improvements are also found in this scenario for the bare category (+40% producer's, +32% user's), Woodlands (+29% producer's), Grasslands (+22% for both user and producer), and Croplands (+25% user's). For both dates, the results obtained for the Shrublands category are rather poor, indicating that this class contains substantial training errors and/or overlaps with other classes in spectral space. Nonetheless, the results indicate that an approach that uses a combination of scenes may be warranted to exploit the benefits of each scene in order to provide improved land cover discrimination.

The mean errors for the 20 testing blocks when using two dates in the classification were substantially reduced at 16.41% (±0.48%), for an average improvement of about 12.08% (±0.46%)

over the September scene alone, and 13.60% (±0.54%) for the January scene. The use of two dates in the classification reduced the errors made using either scene by nearly half and generally reduced the confusion between all cover types. Using results for one training/testing random combination as a typical example, it is found that the reduction in confusion between only the three forest types accounts for over two thirds of the total improvement over just the September scene, and over ~85% when considering the improvement for all the forest types and the woodlands classes. The improvement seen over the January scene is more evenly distributed among the non-forest cover types, with the Forest/Woodland classes still accounting for ~21% of the improvement and the Grasslands class alone showing an improvement of over 19%. Clearly, the usage of two dates is very important to the accurate classification of the cover types considered here.

Table 2 shows the confusion matrix for the same testing samples discussed above and highlights both the successes and shortcomings of the classifier when using the original training data. On Table 2, the labels of the test data are given in the rows while the classification results are given in the columns. In this particular case, the overall error rate is 15.9%. Closer examination of the matrix reveals that 28.72% of the total errors are caused by confusion between the three forest types, with 22.7% just between mixed and deciduous broadleaf forests. The confusion between the woodlands category and the three forest types accounts for 21.5% of the total errors while the errors for the woodlands class alone account for over 34% of the total error. Table 2 also shows that few errors are found between forest and non-forest types. These results are summarized on Table 2 from user's and producer's accuracies. These show that the user's and/or producers accuracies are greater than 80% for eight out of the 11 types considered but also show the poor success obtained for the shrubland class. These highlight the difficulties in separating fairly similar cover types (e.g. forest

types), or cover types that form a continuum between other cover types (e.g. Woodland/Shrubland/ Grassland).

Table 3 shows the mean user's and producers accuracies for all 20 testing blocks when using both dates for the classification. These results confirm the good results obtained in general for both accuracy and stability for most classes, except Woodlands and Shrublands, and to a lesser extent grasslands. The larger standard deviations for the mean accuracies for these classes are also indicative of the difficulties in classifying these classes from the ETM+ data alone. In spectral space, and even over time, the signatures for these classes tend to overlap with those of classes of lower and higher tree densities, such as is the case of some woodlands being confused with both with forest classes and/or shrubs or grasslands. Additionally, because the training process is also somewhat subjective, it is likely that some training pixels for particular classes are mislabeled.

Assuming that some errors have been made during the training phase, the goal of using a consensus filter is to identify those instances that are the most troublesome and thus to improve the quality of the training data that are used to produce the final product. Decision trees are particularly well suited to this type of approach. The approach of using a consensus filter of 20 trees is also conservative with regards to potentially discarding useful data (Brodley and Friedl 1997).

During this filtering process, 691 original training pixels, or 5.14% of the original training data, were misclassified by all 20 trees. Table 4 shows the number of pixels per class, as well as the percentage of pixels per class from the original training data that were identified during this analysis. In terms of proportions per class, 29.1% of all Shrubland training pixels were identified as mislabeled, while all 20 trees misclassified almost 10% of the Woodlands category. All other classes were below 10%. The magnitude of these proportions are expected based on the results presented in Tables 2 and 3, with classes showing more confusion having more mislabeled pixels than others. It

should also be noted that the two smallest classes in terms of training data (Shrublands and Wetlands) have substantially different proportions of pixels identified as mislabeled, indicating the different performance of the classifier for each of these classes. The 691 training pixels identified here were discarded and the new training set of 12758 pixels used to produce the boosted tree and final vegetation map and using both scenes for the classification.

Table 5 provides an evaluation against 195 validation points collected principally from ground-based sources of the final land cover map produced from two ETM+ scenes using a boosted decision tree and the filtered training data. The overall accuracy of this map is 82.05%, with a Kappa coefficient of agreement (Congalton 1991) of 0.80. The forest/non-forest separation is excellent, with only one sample being classified incorrectly, indicating that the usage of multiple ETM+ scenes can provide very accurate results over the use of a single date, particularly for the different forest types. More confusion is found for the Grasslands class as well as the Croplands class, with some errors also associated with the two sparse vegetation classes (Bare, Urban). Because some croplands fields had been already harvested at the time of the September scene, being left bare or with some senescent vegetation, this source of confusion is not unexpected, as similar issues were encountered by Zhu et al. (2000) during their accuracy assessment of the USGS New York/New Jersey regional land cover product. Likewise, the presence of snow in the January scene may have caused these classes to appear similar in spectral space. It is suggestive that perhaps a scene acquired during peak-greenness (e.g. July) may have allowed these classes to be better separated. Nevertheless, the overall accuracy is above the expected accuracy of 80% for these 11 cover types and compares quite favorably with those results published by Zhu et al. (2000), albeit with a different number of land cover classes. Because the cover types found in the general vicinity of the park are very similar to those found outside the park, we fully expect that these validation results can be extended to the larger area

surrounding the park. Figure 4 shows the final land cover map of Delaware Water Gap NRA and surrounding areas. On this figure the park boundaries are shown in black while the boundaries of all the watersheds draining into the park are shown in red.

V. CONCLUSIONS

The methods developed for this pilot project are flexible and can be easily automated. The decision tree methods and techniques developed for coarse scale land cover applications appear to be extendable to high spatial resolution data. The results and land cover classes are in large part consistent with the protocols set forth by the VMP, and compare well with other published results. Perhaps more importantly in the context of projects such as the VMP, the methods are accessible to a wide variety of users, including park scientists and managers who may not always have a lot of expertise with remote sensing. The results demonstrate the applicability of the approach not only to the entire National Park system but also to high spatial resolution land cover applications. These maps can support a wide variety of park activities while also allowing the park to be placed and studied within a more regional context. Moreover, the maps can be produced routinely and at a fairly low cost, further addressing the more immediate needs of many parks for land cover information.

Because this is a pilot project, several avenues for refinement or improvement in the methods/results have been identified. First, the use of an additional Landsat scene during peak greenness may improve the separation of land cover types such as grasslands, pastures, croplands and bare ground. Second, further refinement of the training data is needed to allow the production of a complete map at the formation level of the NVCS (e.g. Deciduous and Evergreen woodlands and shrublands). Third, improvements and possible separation of the wetlands category into grass-dominated wetlands and more forested wetlands may be warranted. Finally, in order to refine the land

cover categories beyond the formation level, the decision tree can be augmented by using other remotely sensed information or other ancillary data sets such as Digital Elevation Models (DEM) and/or soil type information, for example, in turn providing potential improvements in classification accuracies. This type of information could also be used to potentially separate Hemlock stands from other evergreen needleleaf forest types, as Hemlocks preferentially grow in steep ravines with northern exposures.

ACKNOWLEDGEMENTS

The authors would like to thank Richard Evans of Delaware Water Gap NRA for general assistance and helpful discussions during the initial phases of this project. Larry Hilaire provided support with the croplands GIS coverages of the park. Field data were acquired with the assistance of Lisa Solberg, Sandra Mattfeldt and Melissa Stepek during the training and validation phases of the project.

REFERENCES

Breiman, L., J.H. Friedman, R.A. Olshen, and C. Stone (1984*). Classification and Regression Trees*. Wadsworth, Belmont CA.

Brodley, C.E. and M.A. Friedl (1997). Identifying and eliminating mislabeled training instances. In *Proceedings of the 13th National Conference on Artificial Intelligence*, pp. 799-805, AAAI Press, Portland.

Brown de Colstoun, E.C., W. Yang, R. DeFries, M. Hansen and J. Townshend (2000). *Surface Type Visible/Infrared Imager Radiometer Suite Algorithm Theoretical Basis Document, Version 3.0*. Available from World Wide Web site http://npoesslib.ipo.noaa.gov/atbd_viirs.htm

Congalton, R.G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens. Environ.*, 37:35-46.

DeFries, R.S., M. Hansen, and J. Townshend (1995). Global discrimination of land cover types from metrics derived from AVHRR pathfinder data. *Remote Sens. Environ.*, 54:209-222.

DeFries, R.S., M. Hansen, J.R.G. Townshend, and R. Sohlberg (1998). Global land cover classifications at 8 km spatial resolution: the use of training data derived from Landsat imagery in decision tree classifiers. *Int. J. Remote Sens.*, 19:3141-3168.

DeFries, R.S. and J.C.W. Chan (2000). Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote Sens. Environ.*, 74:503-515.

FGDC (1997), *Vegetation Classification Standard*, FGDC-STD-005, Federal Geographic Data Committee, Reston, VA. 58 pp.

Freund, Y. and R.E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119-139.

Friedl, M.A. and C. Brodley (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sens. Environ.*, 61:399-409.

Friedl, M.A., C.E. Brodley, and A.H. Strahler (1999). Maximizing land cover classification accuracies produced by decision trees at continental to global scales. *IEEE Trans. Geosci. Remote Sens.*, GE-37:969-977.

Hansen, M., R. Dubayah, and R. DeFries (1996), Classification trees: an alternative to traditional land cover classifiers, *Int. J. Remote Sens.* 17:1075-1081.

Hansen, M. C., R. S. DeFries, J. R. G. Townshend, and R. Sohlberg (2000). Global land cover classification at 1 km spatial resolution using a classification tree approach. *Int. J. Remote Sens.* 21: 1331-1364.

Muchoney, D., J. Borak, H. Chi, M. Friedl, S. Gopal, J. Hodges, N. Morrow, and A. Strahler (2000). Application of the MODIS global supervised classification model to vegetation and land cover mapping in central america. *Int. J. Remote Sens.*, 21:1115-1138.

Pax-Lenney, M. and C.E. Woodcock (1997). Monitoring agricultural lands in egypt with multitemporal Landsat TM imagery: How many images are needed? *Remote Sens. Environ.* 59:522-529.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufman Publishers, Inc., San Mateo, CA.

Quinlan, J.R. (1996), Bagging, boosting and C4.5, *In Proceedings of the 13th National Conference on Artificial Intelligence*, pp. 725-730, AAAI Press, Portland.

Quinlan, J.R. (1999). Simplifying decision trees. *Int. J. Human-Computer Studies*, 51:497-510.

Richards, J.A. (1983). *Remote Sensing Digital Image Analysis*, Springer Verlag, London.

Schapire, R.E. (1999). A brief introduction to boosting, *In Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pp. 1-6, AAAI Press, Portland.

Strahler, A.H. et al. (1999). *MODIS Land Cover Product*: *Algorithm Theoretical Basis Document, Version 5.0.* Available from World Wide Web Site: http://spso.gsfc.nasa.gov /atbd /modistables.html.

Swain, P.H. and S.M. Davis (1978). *Remote Sensing: The Quantitative Approach*. McGraw-Hill, N.Y.

Vogelmann, J.E., T.L. Sohl, P.E. Campbell, and D.M. Shaw (1998). Regional land cover characterization using landsat thematic mapper data and ancillary data sources. *Environmental Monitoring and Assessment.* 51:415-428.

Weiss, S.M. and C. A. Kulikowski (1991). *Computer systems that learn*. Morgan Kaufman Publishers, San Mateo.

Wolter, P.T., D.J. Mladenoff, G.E. Host and T.R. Crow (1995). Improved forest classification in the northern lake states using multi-temporal landsat imagery. *Photogramm. Eng. Remote Sens.* 61:1129-1143.

Zhu, Z., L. Yang, S.V. Stehman and R.L. Czaplewski (2000). Accuracy assessment for the U.S. geological survey regional land-cover mapping program: New York and New Jersey region. *Photogramm. Eng. Remote Sens.* 66:1425-1435.

Table 1. General land cover categories considered in this study and their correspondence with NVCS formation classes (FGDC 1997). The number of original training pixels for each category used to train the decision trees is also given. General class names are given for clarity. The reader is referred to FGDC (1997) for exact NVCS formation names and definitions.

Land Cover Category	NVCS Formation Code	Number of Training Samples
1) Evergreen Needleleaf Forest	IA8Nc	2306
2) Broadleaf Deciduous Forest	IB2Na	2068
3) Mixed Forest	IC3Na	2287
4) Woodland	IIA4Nb or IIB2Na	1246
5) Shrubland	IIIA2Na or IIIC2Na	351
6) Grassland	VA5Nd	757
7) Marshes or Riparian Vegetation	Various	363
8) Cropland	Various	996
9) Bare/Sparse Vegetation	VIIA1Na or VIIA2Na	896
10) Urban/Developed	Non-vegetated	752
11) Water Bodies	Non-vegetated	1427
Total		**13449**

Table 2. Typical confusion matrix for one of the decision trees using the two Landsat scenes when applied to one 50% testing block (n=6727). The classification results are shown in the columns and the actual labels of the test data are shown in the rows. This matrix highlights the principal areas of agreement and confusion for the classifier. User and producer accuracies for this testing sample are also shown. Class numbers correspond directly to the classes listed in Table 1.

	1 ENeF	2 DBrF	3 MixF	4 Wdld	5 Shrb	6 Gras	7 Wtld	8 Crop	9 Bare	10 Urbn	11 Watr
1 ENeF	1134	8	8	3	0	0	0	0	0	0	0
2 DBrF	24	830	144	35	0	0	0	0	0	1	0
3 MixF	19	104	960	54	5	0	0	2	0	0	0
4 Wdld	14	70	54	445	15	8	0	12	5	0	0
5 Shrb	1	3	3	56	41	41	0	18	11	2	0
6 Gras	0	0	2	16	14	268	0	50	29	0	0
7 Wtld	0	0	1	2	0	0	133	9	14	0	23
8 Crop	1	1	3	13	9	35	1	426	8	1	0
9 Bare	0	0	2	8	3	15	1	44	368	7	0
10 Urbn	0	0	0	0	1	1	1	7	9	357	0
11 Watr	7	0	0	0	0	0	11	0	0	0	696
Prod.	*98.35*	*80.27*	*83.92*	*71.43*	*23.30*	*70.71*	*73.08*	*85.54*	*82.14*	*94.95*	*97.48*
User	*94.50*	*81.69*	*81.56*	*70.41*	*46.59*	*72.83*	*90.48*	*75.00*	*82.88*	*97.01*	*96.80*

Table 3. Mean per-class producer's and user's accuracies obtained against 50% testing samples. Second row for each cover type is the standard deviation of the accuracies for 20 trees.

	Producer's	**User's**
1) ENeF	98.49	94.12
	0.14	0.27
2) DBrF	81.83	78.62
	1.87	1.36
3) MixF	81.24	83.07
	1.44	1.65
4) Woodland	67.99	66.85
	3.09	3.27
5) Shrubland	20.85	42.28
	6.62	5.85
6) Grassland	74.16	69.19
	3.64	2.70
7) Wetland	83.21	88.04
	4.58	3.45
8) Cropland	78.92	79.91
	3.81	2.69
9) Bare	84.12	85.12
	3.11	2.65
10) Urban	94.56	96.05
	1.53	1.66
11) Water	97.13	96.91
	0.74	0.63

Table 4. Number of training pixels per land cover class that have been filtered by consensus filtering. The proportion of pixels filtered with regards to the original data are also given.

	# of Pixels Filtered	**% of Original Training Samples**
1) ENeF	22	0.95
2) DBrF	163	7.88
3) MixF	157	6.86
4) Woodland	124	9.95
5) Shrubland	102	29.06
6) Grassland	40	5.28
7)Wetland	11	3.03
8) Cropland	33	3.31
9) Bare	19	2.12
10) Urban	9	1.20
11) Water	11	0.77

Table 5. Confusion matrix when comparing results from the final boosted tree using two dates of ETM+ data with 195 samples obtained during validation field visits. Here the reference data are in the rows and the classification results in the columns. The overall accuracy is 82.05%. The Kappa coefficient of agreement is 0.80.

	1 ENeF	2 DBrF	3 MixF	4 Wdld	5 Shrb	6 Gras	7 Wtld	8 Crop	9 Bare	10 Urbn	11 Watr	**Total**
1 ENeF	20	0	0	0	0	0	0	0	0	0	0	20
2 DBrF	0	18	3	0	0	0	0	0	0	0	0	21
3 MixF	1	1	12	0	0	0	0	0	0	0	0	14
4 Wdld	0	1	0	12	0	1	1	1	0	0	0	16
5 Shrb	0	0	0	1	8	1	0	0	1	0	0	11
6 Gras	0	0	0	2	1	29	1	3	2	2	0	40
7 Wtld	0	0	0	1	0	0	4	0	0	1	0	6
8 Crop	0	0	0	0	0	6	0	13	3	4	0	21
9 Bare	0	0	0	0	0	0	0	0	18	1	0	19
10 Urbn	0	0	0	0	0	0	0	1	0	13	0	14
11 Watr	0	0	0	0	0	0	0	0	0	0	13	13
Total	21	20	15	16	9	32	6	18	24	21	13	195
User	*95.24*	*90.00*	*80.00*	*75.00*	*88.89*	*90.63*	*66.67*	*72.22*	*75.00*	*61.90*	*100*	
Prod.	*100*	*90.00*	*85.71*	*75.00*	*72.73*	*72.50*	*66.67*	*61.90*	*94.74*	*92.86*	*100*	

LIST OF FIGURES

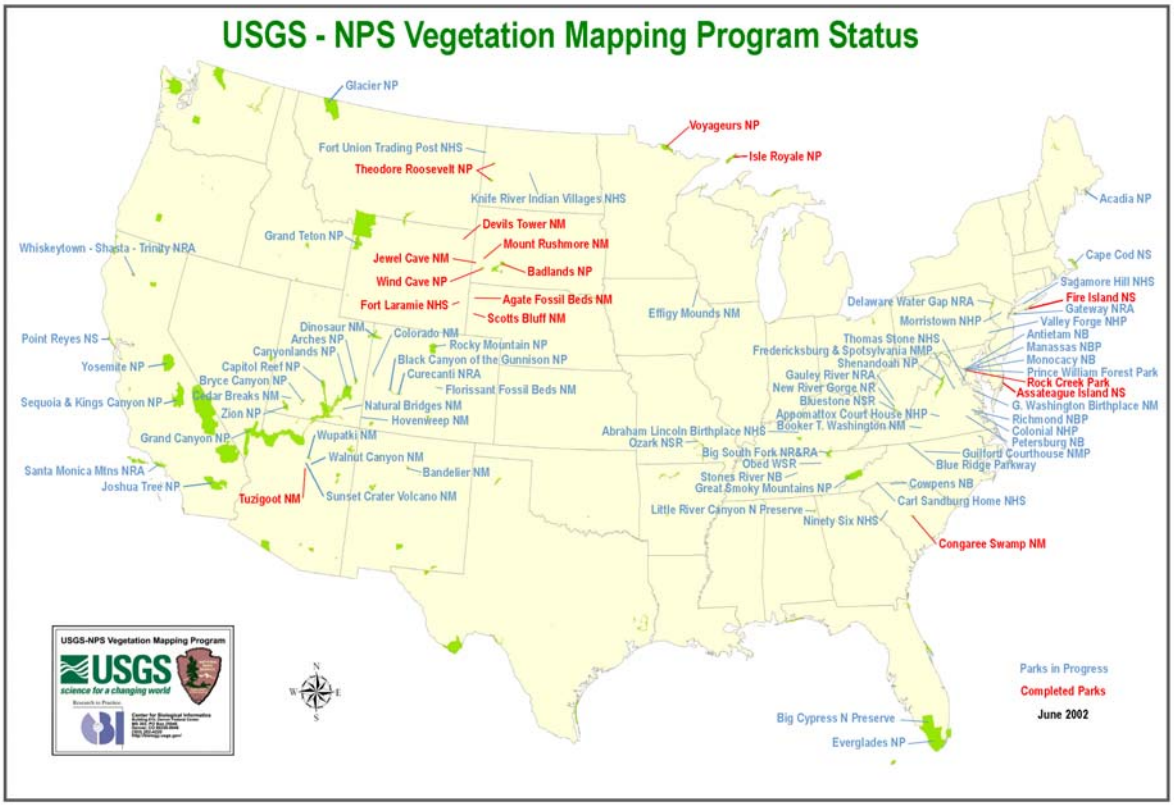
USGS - NPS Vegetation Mapping Program Status
Glacier NP
Voyageurs NP
Fort Union Trading Post NHS
Isle Royale NP
Theodore Roosevelt NP
Knife River Indian Villages NHS
Acadia NP
Devils Tower NM
Grand Teton NP
Mount Rushmore NM
Whiskeytown - Shasta - Trinity NRA
Jewel Cave NM
Cape Cod NS
Badlands NP
Wind Cave NP
Sagamore Hill NHS
Fort Laramie NHS
Agate Fossil Beds NM
Fire Island NS
Effigy Mounds NM
Delaware Water Gap NRA
Gateway NRA
Scotts Bluff NM
Morristown NHP
Dinosaur NM
Valley Forge NHP
Point Reyes NS
Arches NP
Colorado NM
Thomas Stone NHS
Antietam NB
Rocky Mountain NP
Canyonlands NP
Fredericksburg & Spotsylvania NMP
Manassas NBP
Yosemite NP
Capitol Reef NP
Black Canyon of the Gunnison NP
Shenandoah NP
Monocacy NB
Bryce Canyon NP
Curecanti NRA
Gauley River NRA
Prince William Forest Park
Sequoia & Kings Canyon NP
Cedar Breaks NM
Florissant Fossil Beds NM
New River Gorge NR
Rock Creek Park
Natural Bridges NM
Bluestone NSR
Assateague Island NS
Zion NP
Hovenweep NM
Appomattox Court House NHP
G. Washington Birthplace NM
Grand Canyon NP
Wupatki NM
Booker T. Washington NM
Richmond NBP
Abraham Lincoln Birthplace NHS
Colonial NHP
Ozark NSR
Petersburg NB
Walnut Canyon NM
Big South Fork NR&RA
Guilford Courthouse NMP
Santa Monica Mtns NRA
Bandelier NM
Obed WSR
Blue Ridge Parkway
Joshua Tree NP
Stones River NB
Sunset Crater Volcano NM
Great Smoky Mountains NP
Cowpens NB
Tuzigoot NM
Carl Sandburg Home NHS
Little River Canyon N Preserve
Ninety Six NHS
Congaree Swamp NM
USGS-NPS Vegetation Mapping Program
USGS
science for a changing world
N
W
E
S
Parks in Progress
Completed Parks
June 2002
Big Cypress N Preserve
Everglades NP

Figure 1

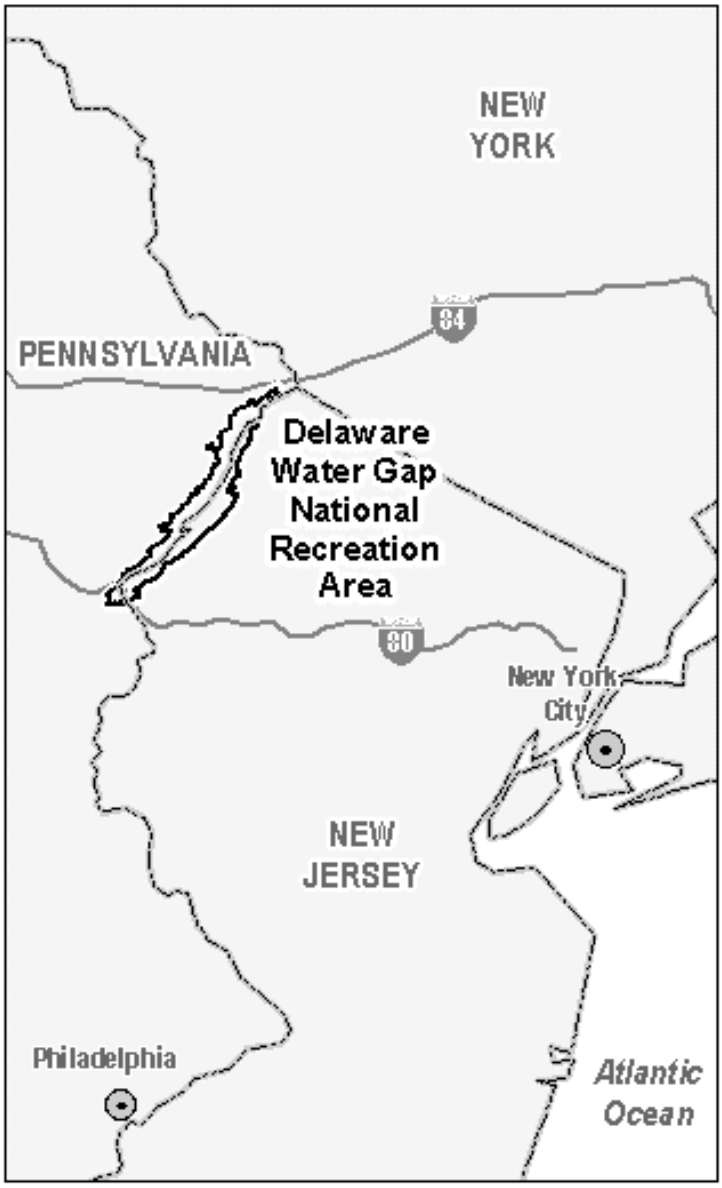
NEW
YORK
84
PENNSYLVANIA
Delaware
Water Gap
National
Recreation
Area
80
New York
City
NEW
JERSEY
Philadelphia
Atlantic
Ocean

Figure 2

Figure 3

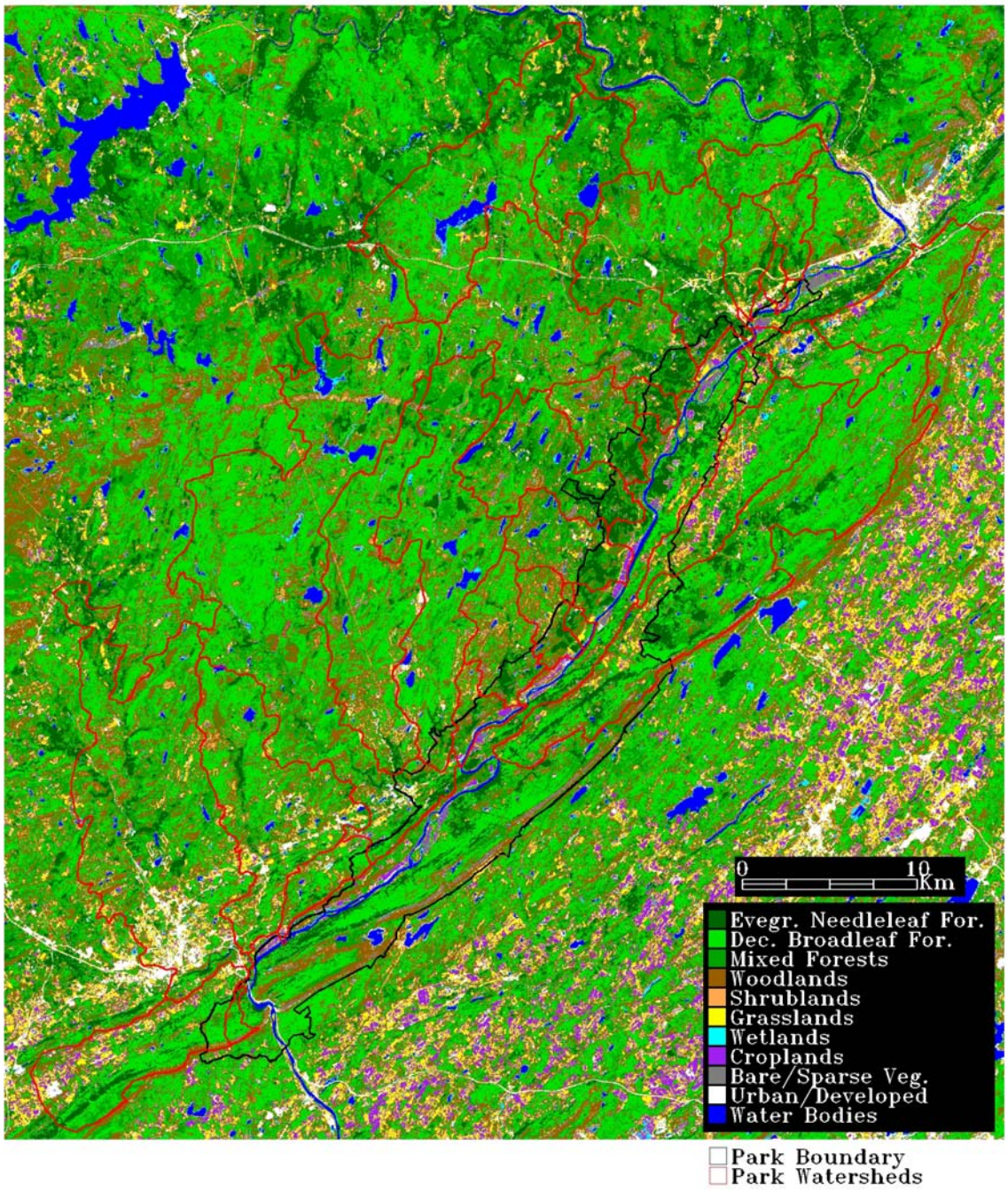
0
10
Km
Evegr. Needleleaf For.
Dec. Broadleaf For.
Mixed Forests
Woodlands
Shrublands
Grasslands
Wetlands
Croplands
Bare/Sparse Veg.
Urban/Developed
Water Bodies
Park Boundary
Park Watersheds

Figure 4